G. Anusha* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.4, Issue No.4, June – July 2016, 3397 – 3400.

Phase Estimation for Single and Multi-Channel Speech Enhancement in Multi Source Environment

G.ANUSHA
M.Tech Student
Department of ECE
Vaagdevi College of Engineering
Warangal, Telangana, India.

M. SHASHIDHAR
Associate Professor
Department of ECE
Vaagdevi College of Engineering
Warangal, Telangana, India.

Abstract: **This letter presents a novel method to estimate the clean speech phase spectrum, given the noisy speech observation in single-channel speech enhancement. The proposed method relies on the phase decomposition of the instantaneous noisy phase spectrum followed by temporal smoothing in order to reduce the large variance of noisy phase, and consequently reconstructs an enhanced instantaneous phase spectrum for signal reconstruction. Multi channel systems utilize spatial diversity which is not present in single channel systems. Novel beam-forming based spatial spectrum estimation methods for multi channel speech enhancement have been proposed in this thesis. Under the fixed beam forming framework, a new reverberant speech enhancement method that utilizes the LP residual cepstrum is developed. On the other hand, a LCMV based spectral method is developed for joint noise cancellation and dereverberation in a beam-forming framework. This is realized as a multi channel LCMV filter that constrains both the early and late parts of the speech frame. The filter outputs are then beam formed to remove late reverberations. These methods indicate significant improvement in perceptual quality of separated signals and distant speech recognition performance when compared to conventional methods.**

Keywords: **Phase Decomposition; Phase Estimation; Speech Enhancement; Speech Quality; Temporal Smoothing.**

I. INTRODUCTION

The rapidly growing market for speech communication systems has been the prime motivation for this thesis. In general, the speech communication systems can be categorized into hands free communication systems, voice controlled systems and hearing aids. Hands free communication systems are widely used in scenarios where limited use of hands is desired. Such scenarios can be hands free car driving and personal navigation systems, where Bluetooth is typically used for communication. Voice controlled systems are used in operation theater by doctors and nurses to move freely around the patients. Hearing aids are typically used by the wearer to amplify the sound to make speech more intelligible. In all the above speech communication systems, the speech source is at a considerable distance from the microphone in a room. The microphone is assumed to be ideal in this thesis, where electrical output is equivalent to the local sound pressure.

Recent speech applications a speech enhancement pre-processor is required to increase the robustness of the overall system against background noise. To this end, previous methods mainly focus on deriving estimators of the clean speech spectral amplitude given the noisy speech while the noisy phase has been typically directly employed for reconstruction of the enhanced signal. The lower branch in Fig. 1 shows the block diagram for the conventional speech enhancement composed of an amplitude modification stage followed by a synthesis stage where the noisy phase spectrum is typically used unchanged to reconstruct the enhanced signal. Many different noise-suppression rules have been proposed to filter the noisy spectral amplitude. The suppression rules are functions of *a priori* and *a posteriori* SNRs estimated from spectral amplitude and noise power spectral density [1]. These methods are either data-driven where training data is exploited as prior knowledge (environment, or user optimized) [2]–[4], or are based on a more general prior knowledge related to probability density functions [5]–[7]. In both groups the noisy phase has been typically utilized in signal reconstruction (for detailed overviews on single-channel speech enhancement see [1] and for an overview on phase importance in speech enhancement see [8]).

The issue of estimating a clean speech phase spectrum has been largely neglected in single-channel speech enhancement. The difficulty in estimating the clean phase spectrum from the noisy signal lies in the fact that the instantaneous phase spectrum is known to jump due to wrapping. Furthermore, no additivity holds to relate the clean speech phase to the noise corrupted phase. Also, early studies reported the unimportance of phase spectrum in perception [9], [10]. This viewpoint

also lies in the fact that the noisy phase was shown to provide the optimal minimum mean square error (MMSE) estimate of the clean phase once the underlying short-time Fourier transform(STFT) coefficients are assumed independent (as was shown for Gaussian speech model [11] and for other speech amplitude distributions [7]). In [7], phase was shown to follow a uniform distribution and to be independent of amplitude when the histogram is calculated from the STFT bins of similar SNR values. This reduces the estimation error variance of the noisy phase. We evaluate the effectiveness of the proposed phase estimation method for two scenarios: directly on the noisy speech, and as a post-processor combined with an amplitude enhancement scheme. Consistent improvement in both perceived quality and intelligibility is achieved compared to when noisy phase is used.

II. SYSTEM DESIGN MODEL

A. *Classification of Speech Enhancement Methods*

The classification of speech enhancement methods is discussed in this section. It is generally difficult for a particular algorithm to perform homogeneously across all types of distortions. Hence, certain assumptions and constraints are required for speech enhancement methods which are generally dependent on specific application and on the environment where it is used.

In general, there are many factors on which the performance of a speech enhancement algorithm is dependent. One of the factors could be number of interfering sources in the multi source environment. In addition to this, assuming different a priori information about the signal of interest or the corrupting signal can also affect the performance of enhancement algorithm. The other factor is the limitation in time variations allowed for the corrupting signal. The last factor is the model based limitation like the restriction of the algorithm to uncorrelated noise. In general, the speech enhancement can be classified in a number of ways. One way to classify speech enhancement methods can be based on single and multiple input channels. They can also be classified based on time and frequency domain processing. The third and last way of classification can be based on adaptive and non adaptive type of algorithms.

In this thesis, the classification based on the number of input channels is used. The brief overview of single and multi channel speech enhancement based classification are explained in the ensuing section.

1. Single Channel Speech Enhancement

In most real time speech based applications, generally a second channel is not available. Such systems are easy to build due to less hardware requirements. Moreover, these single channel systems are comparatively less expensive than the multiple input systems. In the context of noise cancellation, the single channel system constitutes most difficult situations of speech enhancement. In such case, no reference signal to the noise is available and the clean speech cannot be pre-processed prior to being affected by the noise. There are several single channel speech enhancement methods available in the literature such as Wiener filtering, spectral subtraction and cepstral inverse filtering. Such single channel systems utilize different statistics of speech and noise. These systems also assume that noise is stationary during speech intervals. Thus, the performance of single channel methods drastically degrades at lower signal to noise ratios.

2. Multi Channel Speech Enhancement

Single microphone systems only utilize the temporal and spectral diversity of the received signal. Reverberation also induces spatial diversity. To additionally exploit this diversity, multiple microphones should be used. Thus, the beam forming based spatial spectrum estimation techniques have been used in literature for multiple microphone speech enhancement. In the context of noise cancellation, multi channel systems make use of multiple signal inputs to the system and noise reference in an adaptive noise cancellation device. Moreover, the multi channel system utilizes phase alignment to reject undesired noise components. Thus, by exploiting the spatial properties of the signal and the noise source, the non-stationary of noises can be better addressed. This results in overcoming the limitations inherent to one channel systems. The multi channel systems are complex in structure and expensive due to increase in hardware requirement. However, multi channel systems show better speech enhancement results compared to single channel systems.

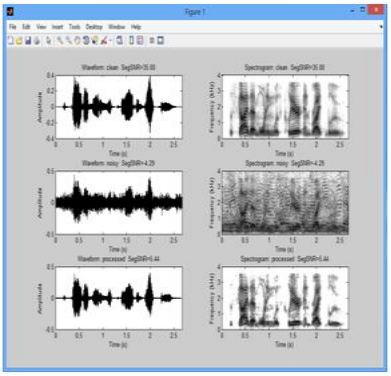
B. *Phase Processing For Speech Enhancement*

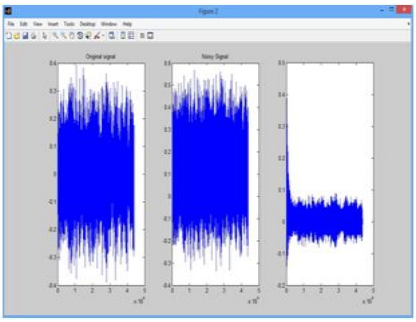
The first proposals for noise reduction in the STFT domain arose in the late 1970s. While the spectral subtraction approaches only modified the spectral magnitudes, the role of the STFT phase was also actively researched at the time. In particular, several authors investigated conditions under which a signal is uniquely specified by only its phase or only its magnitude and proposed iterative algorithms for signal reconstruction from either one or the other . For minimum or maximum phase systems, log-magnitude and phase are related through the Hilbert transform, meaning that only the spectral phase (or only the spectral magnitude) is required to reconstruct the entire signal. But the constraint of purely minimum or maximum phase is too restrictive for real audio signals, and Quatieri

showed that more constraints are needed for mixed-phase signals. For instance, imposing causality or a finite-length constraint on the signal and specifying a few samples of the phase or the signal itself is in some cases sufficient to uniquely characterize the entire phase function from only the magnitude.

The fact that in most state-of-the-art speech enhancement algorithms no phase enhancement is employed, demonstrates that estimating the clean speech phase is a difficult task, and actually a lot more difficult than estimating the amplitude. This has also to do with the fact that the relationship between neighboring phase values in time-frequency space has to be correct. From a statistical point of view, if histograms are computed from STFT-bins that exhibit a similar estimated speech power spectral density, it has been shown that the phase is uniformly distributed and independent of the amplitude [9], [11]. Under these assumptions, it has been shown by Ephraim and Malah, that the MMSE-optimal estimate for the clean speech phases is the noisy phase. This observation tells us that when considering only a certain time-frequency point, the best estimate of the clean speech phase is the noisy phase.

III. SIMULATION RESULTS





For the evaluation of combined phase and amplitude enhancement, a randomly chosen subset of the TIMIT database is deteriorated by additive babble noise at global SNRs ranging from -5 dB to 15 dB in steps of 5 dB. A segment length of 32ms and a segment shift of 4ms is used, at a sampling frequency of 8 kHz. The unbiased MMSE-based noise power estimator proposed in [22] is employed together with the decision-directed approach for the estimation of the a priori SNR [5]. For the estimation of the fundamental frequency, which yields the basis for the phase reconstruction, YIN [23] is used. Compared to [23], the segment shift is adjusted to 4 ms and the threshold for minimum selection is increased to 0.2, which leads to a slightly higher detection rate in low SNR conditions. The subjective, objective and statistical quality evaluation of the separated and dereverberated signals are carried out in this work. The proposed method indicates significant improvements over other conventional methods in literature. Lower word error rates are also noted from distant speech recognition experiments at various DRRs.

IV. CONCLUSION

We presented a new phase enhancement algorithm relying on decomposition of the noisy instantaneous phase and temporal smoothing of the unwrapped phase after removing linear phase. The effectiveness of the proposed method was evaluated in terms of several phase representations showing the harmonic structure in enhanced phase versus the destroyed structure in noisy phase. This method is able to jointly address the problem of noise cancellation, speech dereverberation and speaker separation. The performance evaluation of the proposed method on the GRID corpus indicates that this method is highly robust to noise and reverberation components.

V. REFERENCES

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, “DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement,” in *Synthesis Lectures on Speech and Audio Processing*. San Rafael, CA, USA: . Morgan & Claypool, 2013.

[2] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook-based Bayesian speech enhancement for nonstationary environments,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[3] T. Rosenkranz and H. Puder, “Integrating recursive minimum tracking and codebook-based noise estimation for improved reduction of nonstationary noise,” *Signal Process.*, vol. 92, no. 3, pp. 767–779, 2012.

[4] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, “Cepstral smoothing of spectral filter gains

for speech enhancement without musical noise," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.

[6] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. ICASSP*, Mar. 2008, pp. 4037–4040.

[7] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[8] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," in *Proc. 15th Int. Conf. Spoken Language Processing*, 2014, pp. 1623–1627.

[9] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," in *Proc. IEEE*, May 1981, vol. 69, no. 5, pp. 529–541.

[10] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum- mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[12] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. Int. Conf. Spoken Language Processing*, 2012.

